Modeling competition between vigorousness and dormancy in citation networks

Xue-Wen Wang[1], **Li-Jie Zhang**[1,2], **Guo-Hong Yang**[1,3] and **Xin-Jian Xu**[2,4]

[1]Department of Physics, Shanghai University, Shanghai 200444, China
[2]Institute of Systems Science, Shanghai University, Shanghai 200444, China
[3]The Shanghai Key Lab of Astrophysics, Shanghai, 200234, China
[4]Department of Mathematics, Shanghai University, Shanghai 200444, China

E-mail: wangxuewen@shu.edu.cn, lijzhang@shu.edu.cn, ghyang@mail.shu.edu.cn and xinjxu@yahoo.com

Abstract. In citation networks, the activity of papers usually decreases with age and dormant papers may be discovered and become fashionable again. To model this phenomenon, a competition mechanism is suggested which incorporates two factors: vigorousness and dormancy. Based on this idea, a citation network model is proposed, in which a node has two discrete stage: vigorous and dormant. Vigorous nodes can be deactivated and dormant nodes may be activated and become vigorous. The evolution of the network couples addition of new nodes and state transitions of old ones. Both analytical calculation and numerical simulation show that the degree distribution of nodes in generated networks displays a good right-skewed behaviour. Particularly, scale-free networks are obtained as the deactivated vertex is target selected and exponential networks are realized for the random-selected case. Moreover, the measurement of four real-world citation networks achieves a good agreement with the stochastic model.

Keywords: random graphs, citation networks, degree distribution

1. Introduction

The citation patterns of scientific publications can be simplified into a citation network with nodes representing scientific articles published in journals and edges mimicking citations from one article to another published previously [1]. Citation networks are valuable to uncover the dynamics of scientific publications and have been studied for a long time [2]. A particularly noteworthy contribution was a 1965 study by de Solla Price [3], who proposed the so-called "cumulative advantage"mechanism, that is, a paper which has been cited many times is more likely to be cited again than one which has been little cited. The cumulative advantage is based on the idea of "rich get richer"suggested by Yule [4] and Simon [5], and the criterion now is widely known as the "preferential attachment"in the framework of currently fashionable evolving network models, proposed by Barabási and ALbert in 1999 [6]. By employing growth and preference, the Barabási-Albert (BA) model provides a natural explanation for the scale-free behavior observed in many realistic systems. Recently, Clauset et al. [7] proposed a statistical framework for determining power-law tails of various data sets, in accordance with the conclusion of Redner [8].

In the study of citation networks, one of the most important topics is the characterization of the probability distribution of the number of citations received by a paper and the design of simple microscopic models to reproduce the real-world distribution [9]. Many empirical studies in citation networks have proved that age may be one of the most important mechanisms that determines the statistical properties of the growing network [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21]. To investigate the effect of age on network evolution, the BA model has been modified by incorporating time dependence in citation networks. Dorogovtsev and Mendes DM00 studied the case that the probability of an old node attached by a newcomer is not only proportional to its degree k but also to a power of its age $\tau^{-\alpha}$ (where τ is the age of a node). They found that the resulting network shows scale-free (SF) behavior only in the region $\alpha < 1$. For $\alpha > 1$, the degree distribution $P(k)$ is exponential. One the other hand, Klemm and Eguíluz [11] proposed a degree-dependent deactivation network model, which is highly clustered and retains the power-law distribution of the node's degree.

Most previous studies only consider the irreversible impact of age, such as gradual aging [10] and absolute deactivation [11]. In the real world, however, there is a universal phenomenon called "delayed recognition", that is, papers did not seem to achieve any sort of recognition until some years after their original publication [22, 23]. The question therefore arises as to whether such process can be explained or expected by the network theory. In this paper we express the notion of the delayed recognition in terms of an evolving network model with transitions of nodes' states to answer this question. Intuitively, we suggest that the activity of a node is the result of the competition of two factors: vigorousness and dormancy. For vigorousness, supposing that a new published paper or an old paper, its ability of receiving citations from others increases gradually with time. Whereas for dormancy, it describes the deactivation of the paper and being

slept. The evolution of the network couples addition of new nodes and state transitions of old ones. It is found that the degree distribution of the resulting network depends on the transition probability. Furthermore, we study four real-world citation data and notice the good agreement with present model.

2. Model

The evolution process starts with an initial network of a small number m_0 of isolated nodes, in which m ($m < m_0$) nodes are vigorous. Motivated by previous research [11, 20], at each time step the dynamics runs as follows.

(i) Adding a new node i with m outgoing links that are attached to previously existing m vigorous nodes. We assume that m is the average number of references per article. By $k^{'}$ we denote the in-degree of a node, i.e., the number of edges pointing to it. The in-degree of the newcomer is $k_i^{'} = 0$ at first. Each selected vigorous node j receives exactly one incoming edge, thereby $k_j^{'} \to k_j^{'} + 1$. Since the out-degree of each node is m always, the total degree of a node is $k = k^{'} + m$.

(ii) Activating the new node i, which means that the new published paper is always assumed to be vigorous at first.

(iii) Awakening one of the previously existing dormant nodes. For simplicity, we assume that each dormant node is chosen uniformly to be activated.

(iv) Deactivating two of the vigorous nodes. The probability of a vigorous node j being deactivated is given by

$$\nu(k_j') = \frac{\gamma - 1}{\alpha + k_j'}, \tag{1}$$

where $\alpha > 0$ is a preferential factor reflecting the initial attractiveness of different fields, and the normalization factor is defined as $\gamma - 1 = [\sum_{l\in\Lambda} 1/(\alpha + k_l)]^{-1}$. The summation runs over the set Λ of the currently vigorous nodes. Eq. (1) means that the most cited paper is less possibility to be forgotten.

According the model definition, vigorous nodes may become dormant ones gradually, which can be explained as a collective "forgetting". At the same time, dormant nodes may be awaked and receive links from subsequent node again, which considers the recognition of "forgotten"papers.

3. Degree distribution

Denoting $A_{k'}^t$ the number of vigorous nodes with in-degree k' at time t, one can write out the differential equation

$$\frac{\partial A_{k'+1}^t}{\partial t} = (1-2\nu_{k'}^t)(A_{k'}^t+\mu_{k'}^t)-A_{k'+1}^t = \left(1 - 2\frac{\gamma-1}{\alpha+k'}\right)(A_{k'}^t+\mu_{k'}^t)-A_{k'+1}^t \tag{2}$$

for network evolution, where $\mu_{k'}^t$ is the activation probability. Imposing the stationary condition $\partial A_{k'}^t/\partial t = 0$, one obtains

$$A_{k'+1} - A_{k'} = -2\frac{\gamma-1}{\alpha+k'}A_{k'} + \left(1 - 2\frac{\gamma-1}{\alpha+k'}\right)\mu_{k'}^t. \tag{3}$$

The probability of a dormant node being activated is assumed to be uniform, so $\mu_{k'}^t$ takes the form

$$\mu_{k'}^t = \frac{N_{k'}^t}{m_0 + t - m - 2}, \tag{4}$$

where $N_{k'}^t$ represents the number of dormant nodes with in-degree k' at time t. For large t, the total number of nodes in the network is approximately equal to the number of dormant nodes, and the overall in-degree distribution $n_{k'}$ can be approximated by considering the dormant nodes only. Thus, we obtain the relationship

$$\mu_{k'}^t = n_{k'}, \tag{5}$$

and $n_{k'}$ can be calculated as the rate of the change of vigorous nodes $A_{k'}$,

$$n_{k'} = A_{k'} - A_{k'+1}. \tag{6}$$

Substituting Eqs. (5) and (6) into Eq. (3) yields

$$A_{k'} = A_0 \prod_{i=0}^{k'} \frac{i + \alpha + 2 - 2\gamma}{i + \alpha + 1 - \gamma} = A_0 \exp\left[\sum_{i=0}^{k'} \ln(1 + \frac{1-\gamma}{i + \alpha + 1 - \gamma})\right], \tag{7}$$

where the boundary value A_0 is equal to 1 reflecting the constant addition of newcomers with initial $k' = 0$. In the following, we give analytical solutions corresponding to different α.

(i) The case of samll α and $\alpha > m$. By the approximately logarithmic Taylor expansion, Eq. (7) can be written as

$$A_{k'} = (\alpha + 1 - \gamma)^{\gamma-1}(k' + \alpha + 1 - \gamma)^{-(\gamma-1)}, \tag{8}$$

and the overall in-degree distribution $n_{k'}$ is

$$n_{k'} = -\frac{dA_{k'}}{dk'} = c(k' + \alpha + 1 - \gamma)^{-\gamma}. \tag{9}$$

The normalized factor is $c = (\gamma - 1)(\alpha + 1 - \gamma)^{\gamma-1}$. The exponent γ can be obtained from a self-consistency condition $m = \int_0^\infty k' n_{k'} dk'$, which gives

$$\gamma = 1 + \frac{m + \alpha}{m + 1}. \tag{10}$$

Thus, the exponent γ depends on the parameters α and m. If it is set $\alpha = m + 2$, then one has $\gamma = 3$. Figure 1 shows the total degree distribution obtained by simulating the model for 10^5 time steps. As expected, we obtain power-law distributions with best-fitted exponent γ equal to 2.82(9), 2.92(9), and 2.96(5), corresponding to $m = 10$, 20, and 40, respectively.

(ii) The case of $\alpha \to \infty$. The deactivation probability $\nu_{k'}$ is independent of k', which means that each of the $m + 2$ vigorous nodes will be deactivated with the same probability $1/(m + 2)$. Thus, Eq. (7) can be written as

$$A_{k'} = \exp\left[k' \ln \frac{\alpha + 2 - 2\gamma}{\alpha + 1 - \gamma}\right] = \left(\frac{m}{m+1}\right)^{k'}. \tag{11}$$

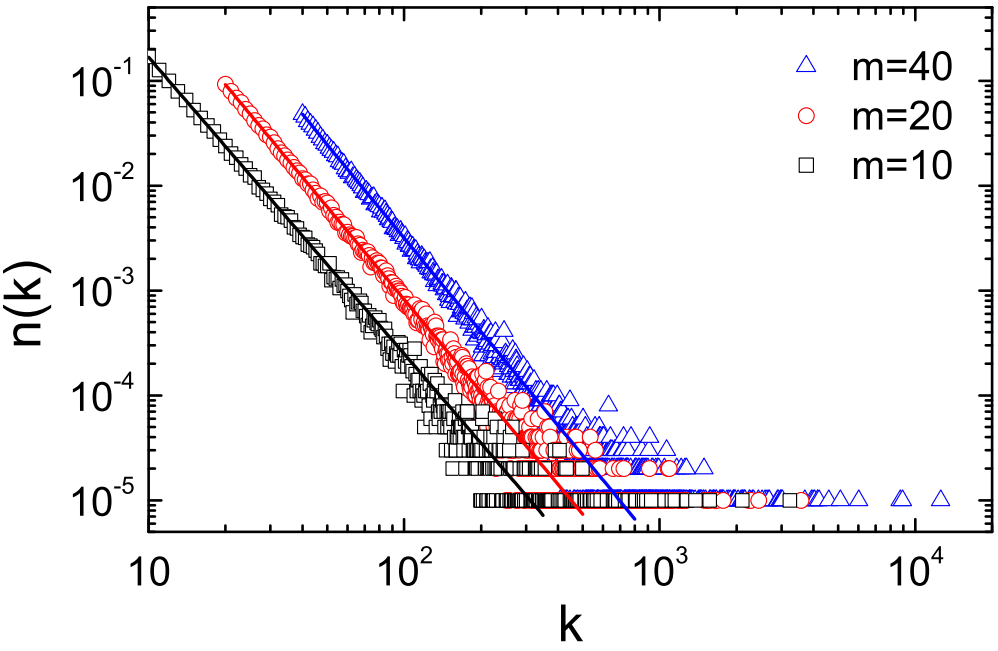

Figure 1. (Color online) Degree distributions of nodes of generated networks in case of $\alpha = m + 2$ for $m = 10$ (square), 20 (circle) and 40 (triangle), respectively. The size of networks is $N = 10^5$. The solid lines are least-squares fits based on the form of Eq. (9).

Then, the overall in-degree distribution $n_{k'}$ is

$$n_{k'} = -\frac{dA_{k'}}{dk'} = \ln\left(\frac{m+1}{m}\right)\left(\frac{m}{m+1}\right)^{k'}. \tag{12}$$

To obtain the total degree distribution, we rewrite the above equation as

$$n_k = \ln\left(\frac{m+1}{m}\right)\left(\frac{m}{m+1}\right)^{k-m}, \tag{13}$$

where $k = k' + m$. Thus, the distribution is exponent decay. In Fig. 2 we plot the total degree distribution of the simulated networks for $m = 10$, 20, and 40, respectively. As expected, we obtain exponential distributions with best-fitted exponent $m/(m+1)$ being 0.90(9), 0.95(2), and 0.97(5), corresponding to $m = 10$, 20, and 40, respectively.

(iii) The case of $m \ll \alpha < \infty$. As k' is small, $A_{k'}$ can be approximated by Eq. (11). While k' is large, $A_{k'}$ can be described by the approximately logarithmic Taylor expansion. Therefore, there exists a tipping point k_c in the degree distribution. As k' is smaller than k_c, Eq. (7) can be simplified to

$$A_{k'} = \left(\frac{m}{m+1}\right)^{k'}. \tag{14}$$

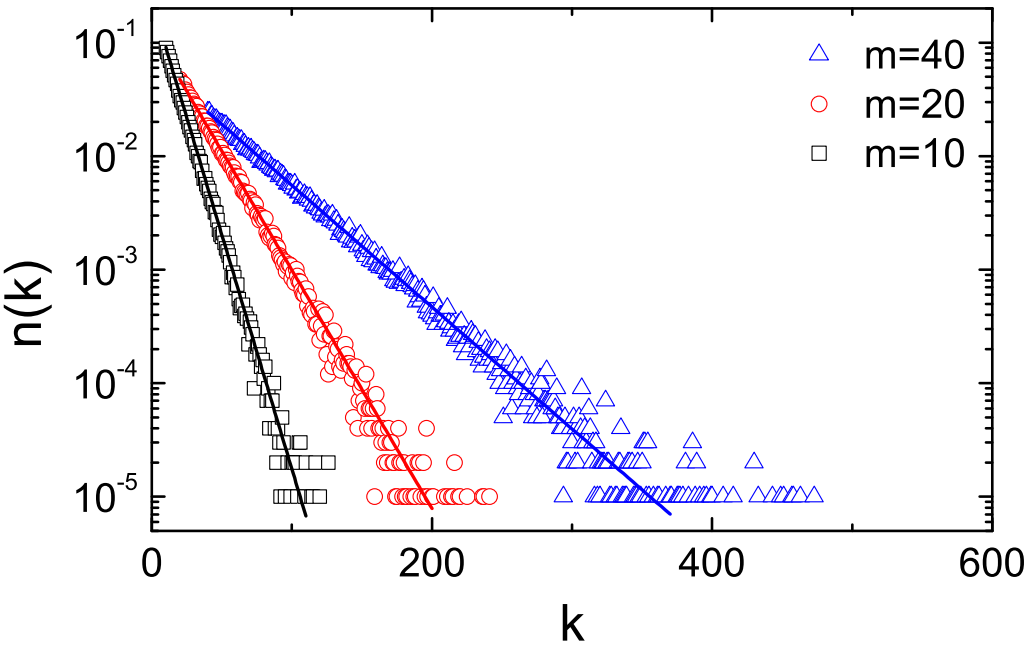

Figure 2. (Color online) Degree distributions of nodes of generated networks in case of $\alpha \to \infty$. The solid lines are least-squares fits based on the form of Eq. (12).

While k' is larger than k_c, Eq. (7) reduces to

$$A_{k'} = \left(\frac{\alpha+2-2\gamma}{\alpha+1-\gamma}\right)^{k_c} (k_c+\alpha+1-\gamma)^{\gamma-1} \times (k'+\alpha+1-\gamma)^{-(\gamma-1)}. \quad (15)$$

Combining aove two expressions, one can obtain the overall in-degree distribution $n_{k'}$

$$n_{k'} = (\gamma-1)\left(\frac{\alpha+2-2\gamma}{\alpha+1-\gamma}\right)^{k_c} \times (k_c+\alpha+1-\gamma)^{\gamma-1}(k+\alpha+1-\gamma)^{-\gamma}. \quad (16)$$

In Fig. 3, we plot the total degree distribution of the generated networks with parameters $\alpha = 200$ for $m = 10$, 20, and 40, respectively. All the plots are right-skewed, in agreement with the theoretical prediction.

4. Comparison with empirical data

To examine present model, we utilize four empirical data from citation networks.

(i) PNAS data [24], which contains 23, 572 articles and 40, 853 edges published by the proceedings of the National Academy of Sciences (PNAS) of the United States of America from 1998 to 2007.

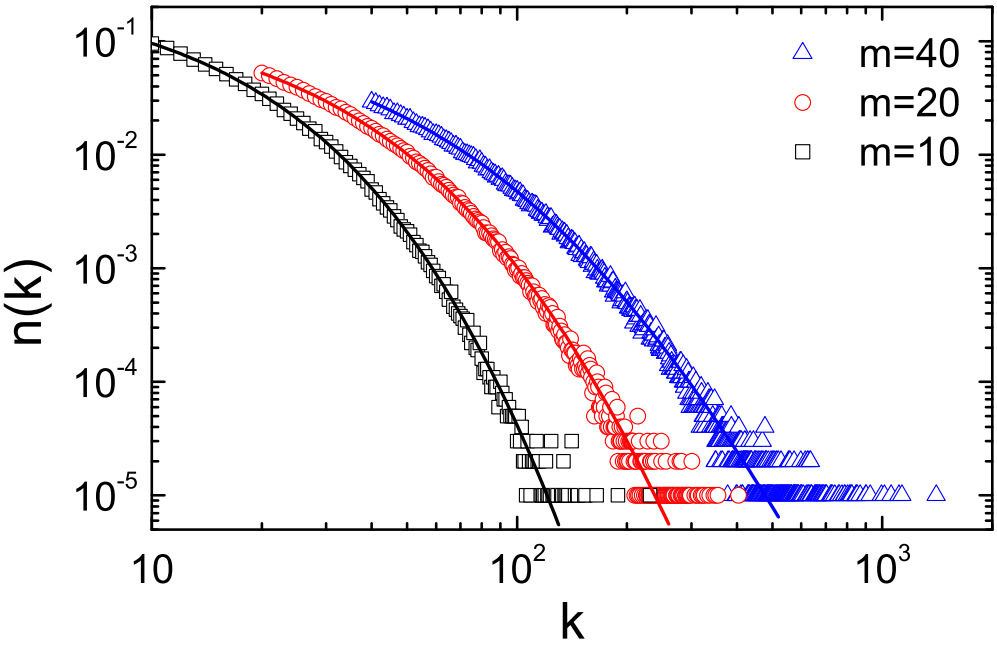

Figure 3. (Color online) Degree distributions of nodes of generated networks in case of $m \ll \alpha < \infty$. The solid lines are least-squares fits based on the form of Eq. (16).

(ii) Hep-th data [25], which comes from preprints posted on arxiv.org, and covers papers in the period from January 1992 to April 2003 (124 months). It contains 27, 770 papers and 352, 807 edges.

(iii) Hep-ph data [25], which comes from preprints posted on arxiv.org, and covers papers in the period from January 1992 to April 2003 (124 months). It contains 34, 546 papers and 421, 578 edges.

(iv) U.S. Patent data [26], which is maintained by the National Bureau of Economic Research. The data includes all citations made by patents granted between 1975 and 1999, and contains 3, 774, 768 nodes and 16, 518, 948 edges.

Figure 4 shows the comparison of degree statistics of four citation networks with numerical results of generated networks. To gain values of m_0, m and α, which refer to the number of initial isolate nodes, the number of references per paper and the attractiveness bias, respectively, we fit the empirical distribution based on Eq. (16). Although the empirical networks are different in nature, all the cumulative in-degree distribution follow a right-skewed decay, which shifts from an exponential to a power law. Table 4 shows empirical data on the citation distribution of papers and assessed parameters m_0, m and α by simulation, and one notices the good agreement.

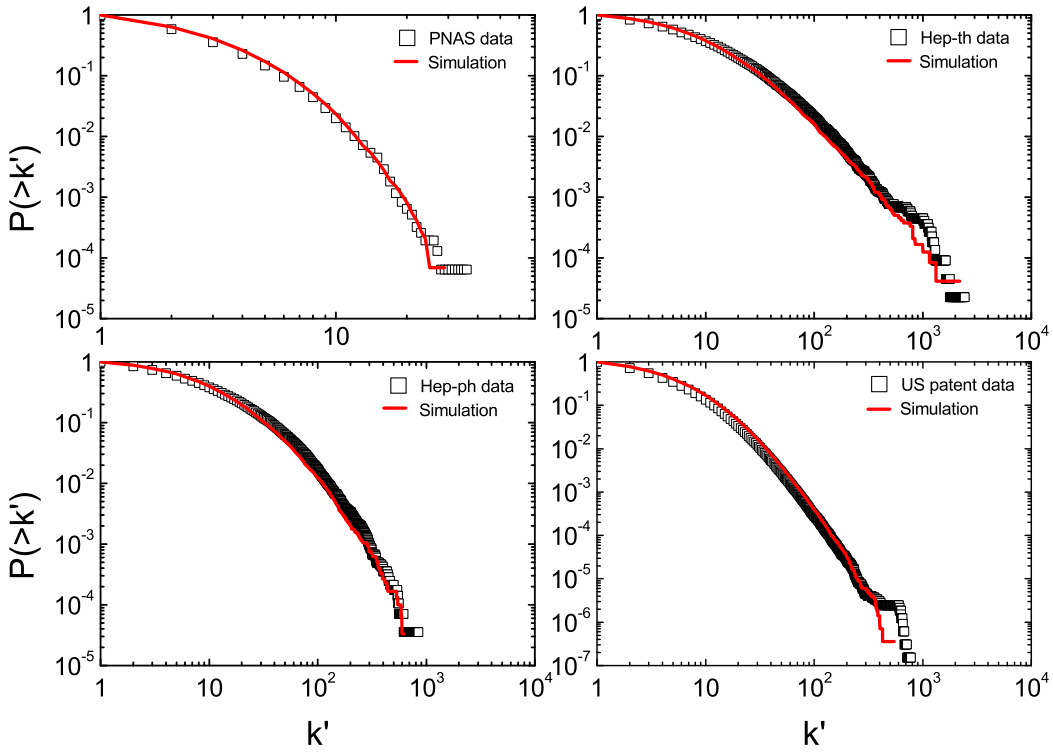

Figure 4. (Color online) Comparison of empirical networks with simulation results of the present network model. Parameters of simulations for different values $m_0 = 3145$, $m = 2$, $\alpha = 24$; $m_0 = 631$, $m = 13$, $\alpha = 12$; $m_0 = 2117$, $m = 13$, $\alpha = 18$; $m_0 = 470,978$, $m = 5$, $\alpha = 12.5$ correspond to PNAS, Hep-th, Hep-ph and U.S. Patent data, respectively.

5. Conclusion

In summary, we have proposed a simple model for citation networks to explain the phenomenon of delayed recognition in the life of a article which usually begins lesser, rises to peak, and then diminishes. We suggested that the activity of a paper is the result of the competition of vigorousness and dormancy. The growth dynamics of the network is governed by the state transition. We found that the average number of references per paper m and the initial attractiveness of different fields α determine the topological structure of the generated network. If the value of α is selected appropriately as $m + 2$, the deactivation probability $\nu(k)$ is a linear preferential one, which leads to a power-law degree distribution with the exponent $\gamma = 3$. Whereas for α tends to ∞, the vigorous nodes are selected to be deactivated with the uniform probability, and the model gives rise to an exponential degree distribution with the exponent only depending on m. Between the two regimes, the distribution gradually shifts from the exponential to the power law. To examine theoretical prediction, we compared the degree distribution

Table 1. Basic statistics of PNAS, Hep-th, Hep-ph and U.S. Patent data. N, E and $\overline{k}$ denote the number of nodes, edges and average out-degree in four empirical networks, respectively. N', E', m_0, m and α are parameters for simulated networks. N' and E' denote the number of nodes and edges of the networks. m_0 represent the initial isolated nodes. m and α represent the average out-degree and the constant bias in the networks.

$\frac{Measures}{networks}$	PNAS	Hep-th	Hep-ph	U. S. Patent
N	23,572	27,770	34546	3,774,768
E	40853	352,807	421578	16,518,948
$\overline{k}$	1.7	12.7	12.2	4.4
N'	23,572	27,770	34546	3,774,768
E'	40853	352,807	421578	16,518,948
m_0	3145	631	2117	470,978
m	2	13	13	5
α	24	12	18	12.5

with empirical citation networks and noticed a good agreement. So the present model provides a new way to understand citation networks with age.

Acknowledgments

This work was supported by the Innovation Program of Shanghai Municipal Education Commission (13YZ007) and the Specialized Research Fund for the Doctoral Program of Higher Education under No. 20093108110004.

References

[1] Redner S, 2005 *Phys. Today* **58** 49
[2] Egghe L and Rousseau R, 1990 *Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science* (New York, Elsevier)
[3] de Solla Price D J, 1965 *Science* **149** 510
[4] Yule G U, 1925 *Phil. Trans. R. Soc.* B **213** 21
[5] Simon H A, 1955 *Biometrika* **42** 425
[6] Barabási A L and Albert R, 1999 *Science* **286** 509
[7] Clauset A, Shalizi C R and Newman M E J, 2009 *SIAM Rev.* **51** 661
[8] Redner S, 1998 *Eur. Phys. J.* B **4** 131
[9] Radicchi F, Fortunato S and Castellano C, 2008 *Proc. Natl. Acad. Sci. U.S.A.* **105** 17268
[10] Dorogovtsev S N and Mendes J F F, 2000 *Phys. Rev.* E **62** 1842
[11] Klemm K and Eguíluz V M, 2002 *Phys. Rev.* E **65** 036123
[12] Zhu H, Wang X R and Zhu J Y, 2003 *Phys. Rev.* E **68** 056121
[13] Vázquez A et al, 2003 *Phys. Rev.* E **67** 046111
[14] Hajra K B and Sen P, 2006 *Physica* A **368** 575

[15] Tian L et al, 2006 *Phys. Rev.* E **74** 046103
[16] Wang M, Yu G and Yu D, 2008 *Physica* A **387** 4692
[17] Gingras Y et al, 2008 *PLoS ONE* **3** e4048
[18] Crokidakis N and de Menezes M A 2009 *J. Stat. Mech.* P04018
[19] Xu X J and Zhou M C, 2009 *Phys. Rev.* E **80** 066105
[20] Xiong F et al, 2011 *Eur. Phys. J.* B **84** 115
[21] Golosovsky M and Solomon S, 2012 *Phys. Rev. Lett.* **109** 098701
[22] Van Raan A F J, 2004 *Scientometrics* **59** 467
[23] Burrell Q L, 2005 *Scientometrics* **65** 381
[24] Ren F X, Shen H W and Cheng X Q, 2012 *Physica* A **391** 3533
[25] Gehrke J, Ginsparg P and Kleinberg J, 2003 *ACM SIGKDD Explorations Newsletter* **5** 149
[26] Leskovec J, Kleinberg J and Faloutsos C, 2005 *Proceedings of the Eleventh ACM SIGKDD International Conference on knowledge Discovery and Data Mining* (ACM, New York, USA) pp. 177-187